

## Прежде всего сформулируем условия задачи:

1. Выбрать ген - выбран **Ribosomal protein S10**. Далее выбрать виды, в геноме которых есть данный ген. Минимальное число видов, необходимых для включения в дерево – 10. Больше число видов 15-20 приветствуется. Найти аминокислотные последовательности генов для нижеперечисленных видов и составить fasta-файл с последовательностями. Поставить в начало названия гена строку с названием организма, чтобы организм отображался на листьях дерева. Например:

Bacterium ANI16583.1 ATP synthase alpha [Lactobacillus sp. HRR0T3]

Хорошо бы иметь представленность следующих видов:

Человек - human

Обезьяна - primates

Грызуны – mouse, rat

копытное - bovine

сумчатое - marsupials

пресмыкающиеся – snakes, lizards, turtles

птица

рыба

растение

fungi (*Saccharomyces cerevisiae*) – пекарские дрожжи (продаются сухими в пакетиках)

архея - archaea

бактерия – bacteria

Поиск гена производится на сайте NCBI в базе данных “protein”. Набиваете название гена в скобках. (ribosomal protein L1) + используйте кнопку Advanced для лучшего поиска.

2. Произвести множественное выравнивание последовательностей алгоритмами ClustalW и Muscle из пакета Mega. Сохранить два выравнивания в отдельные файлы. Написать, есть ли разница в выравнивании.
3. В программе Mega построить филогенетическое дерево для аминокислотных последовательностей с бутстрэп-анализом для двух выравниваний – и методом ClustalW, и Muscle.
  - методом расстояний (UPGMA)

- методом расстояний (NJ)
- методом максимального правдоподобия

В качестве отчетности представить файл в формате fasta с аминокислотными последовательностями файлы с двумя выравниваниями – clustalw и muscle отчет, оформленный в виде pdf или ppt: скриншот выравнивания для метода ClustalW и Muscle скриншоты деревьев, построенных (1) методом расстояний (UPGMA), (2) методом расстояний (NJ), методами максимального правдоподобия, на которых будут видны бутстрэп-значения, а также хорошо читаемые названия организмов. Выводы Какой алгоритм выравнивания лучше сработал - ClustalW или Muscle? Одинаковая ли получилась топология деревьев при построении разными методами? Одинаковые ли получились бутстрэп-значения? Совпадают ли деревья, построенные по одному гену с принятыми деревьями видов?

## Скачивание данных:

Для начала выберем представителей для каждой из запрашиваемых категорий:

1. Человек - *Homo sapiens*
2. Обезьяна - *Macaca fascicularis*
3. Крыса - *Rattus norvegicus*
4. Мышь - *Mus musculus*
5. Копытное - *Bos taurus*
6. Сумчатое - *Phascogale carolinensis*
7. Змея - *Protobothrops mucrosquamatus*
8. Ящерица - *Podarcis muralis*
9. Черепаха - *Chelonia mydas*
10. Птица - *Columba livia*
11. Рыба - *Takifugu rubripes*
12. Растение - *Zea mays*
13. Пекарские дрожжи - *Saccharomyces cerevisiae*
14. Архея - *Pyrococcus woesei*
15. Бактерия - *Pseudomonas koreensis*
16. Ракообразные - *Ancylostoma caninum*
17. Кролик - *Oryctolagus cuniculus*
18. Кукурузные - *Trichosurus vulpecula*

После чего полученные последовательности объединяем в один .fa файл:

ФайлПравкаПоискВидКодировкиСинтаксисыОпцииИнструментыМакросыЗапускПлаг

RibProt\_S10.fa

45

>NP\_001105007.1 ribosomal protein S10 [Zea mays]

46

MAAKIRVVMKSFMSQSNQVVGLLPFTKKVGLPESRALYTVLRSPHIDKKSREQFSMHVKKQFVELTAKPH

47

ELHKFFWLKRLRIPGAQYEVQISFKTRLDMASLRSQAP

48

>KAF4006697.1 ribosomal protein S10 [Saccharomyces cerevisiae]

49

MLRNTIALRSFIRTQSTRPYPVNVEAVYYAPLKLPIKYGDLVADIQLRSYDNENLDFYSDFILRTGYL

50

IPLTGPKPLPTRRERWTVIKSPFVHAWSKENFERHTHKRLIRAWDTNPEVLQMLIAYITKHSMAVGVMKC

51

NFFQRSEISLDLGSDANGLEKSLSNIDELYSLRNDDKAQTSAVGQKVLELLDSPDFKKHLEKK

Выравнивание:

Для начала рассмотрим результат:

Protein Sequences	
Species/Abbrv	*
1. AAA85660.1 ribosomal protein S10 Homo sapiens	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
2. XP_005553317.1 40S ribosomal protein S10 Macaca fascicularis	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
3. AAH58141.1 Ribosomal protein S10 Rattus norvegicus	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
4. AAH89323.1 Ribosomal protein S10 Mus musculus	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
5. AAI02417.1 Ribosomal protein S10 Bos taurus	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
6. XP_020855249.1 40S ribosomal protein S10 Phascolarctos cinereus	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
7. XP_015675910.1 40S ribosomal protein S10 Protobothrops mucrosquamatus	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
8. XP_028589910.1 40S ribosomal protein S10 Podarcis muralis	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
9. XP_007063566.1 40S ribosomal protein S10 Chelonia mydas	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
10. PKK19037.1 ribosomal protein S10 Columba livia	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L V D K N
11. XP_029682140.1 40S ribosomal protein S10 isoform X1 Takifugu rubripes	M F M M L M P K K N R I A I Y E L L F K E G V M V A K K D V H L T K H P E L A
12. NP_001105007.1 ribosomal protein S10 Zea mays	M A A K I R V V M K S F M S Q S N Q V V G L L P F T K K V G L P E S R A L Y T V
13. KAF4006697.1 ribosomal protein S10 Saccharomyces cerevisiae	M L R N T I A L R S F I R T Q S T R P Y P V N V E A V Y Y A P L K L P I K Y G
14. CAA42518.1 ribosomal protein S10 Pyrococcus woesei	M Q K A R I K I A S T N V R S L D E V A N Q I K Q I A E R T G V R M S G P I P
15. RVD76133.1 ribosomal protein S10 Pseudomonas koreensis	M Q N Q Q I R I R L K A F D H R L I D Q S T Q E I V E T A K R T G A Q V R G P
16. RCN50307.1 ribosomal protein S10 Ancylostoma caninum	M R I T Y R S H F A V C A Q L I E G A K N E N L V V K G P I R L P T K V L R I
17. XP_002718159.1 40S ribosomal protein S10 Oryctolagus cuniculus	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N
18. XP_036591903.1 40S ribosomal protein S10 Trichosurus vulpecula	M L M P K K N R I A I Y E L L F K E G V M V A K K D V H M P K H P E L A D K N

Видно, что для разных организмов последовательность одинаковая, что говорит о консервативности гена.

[https://drive.google.com/drive/folders/1xbIFq3xWgH81rSToDC\\_TXKGB96DIZTnc?usp=\\*sharing](https://drive.google.com/drive/folders/1xbIFq3xWgH81rSToDC_TXKGB96DIZTnc?usp=*sharing)\*

**Теперь сделаем выравнивание двумя способами.**

ClustalW:

Protein Sequences	
Species/Abbrv	
1. AAA85660.1 ribosomal protein S10 Homo sapiens	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
2. XP_00553317.1 40S ribosomal protein S10 Macaca fascicularis	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
3. AAH8141.1 Ribosomal protein S10 Rattus norvegicus	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
4. AAH89323.1 Ribosomal protein S10 Mus musculus	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
5. AAI02417.1 Ribosomal protein S10 Bos taurus	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
6. XP_020855249.1 40S ribosomal protein S10 Phasciarctos cinereus	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
7. XP_015675910.1 40S ribosomal protein S10 Protobothrops mucroscquamatus	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
8. XP_028589910.1 40S ribosomal protein S10 Podarcis muralis	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
9. XP_007063566.1 40S ribosomal protein S10 Chelonia mydas	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
10. PKK119037.1 ribosomal protein S10 Columba livia	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
11. XP_029682140.1 40S ribosomal protein S10 isoform X1 Takifugu rubripes	MFMMLMPKKNRITAIYEELLFKEG--VMVAKKDVHMT-KHPELADKNVPLNHVMKAMQSLKSCGYVK--EQFAWRHYYW
12. NP_001105007.1 ribosomal protein S10 Zea mays	-----MLRNTIALRSFIRTSQSTRPVNVVEAVVYAPLKLPL-----MAAKIK
13. KAF4006697.1 ribosomal protein S10 Saccharomyces cerevisiae	-----MLRNTIALRSFIRTSQSTRPVNVVEAVVYAPLKLPL-----IKYGLVDADIK
14. CA442518.1 ribosomal protein S10 Pyrococcus woesei	-----MLRNTIALRSFIRTSQSTRPVNVVEAVVYAPLKLPL-----MQKARI
15. RVD76133.1 ribosomal protein S10 Pseudomonas korensis	-----MLRNTIALRSFIRTSQSTRPVNVVEAVVYAPLKLPL-----MQNQRI
16. RCN50307.1 ribosomal protein S10 Ancylostoma caninum	--MRITYRSHFAVCAQLIEGAKNENLVVKGPIRLPTKVLRIITTRKTPCGEGSKTWDRFGHLLFLKFYCTSNQIMAGIAYKNIEKPLPDNTEHRRIRL
17. XP_002718159.1 40S ribosomal protein S10 Oryctolagus cuniculus	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW
18. XP_036591903.1 40S ribosomal protein S10 Trichosurus vulpecula	--MLMPKKNRITAIYEELLFKEG--VMVAKKDVHMP-KHPELADKNVPLNHVMKAMQSLKSRGYVK--EQFAWRHFFYW

**MUSCLE:**

Protein Sequences	
Species/Abbrv	
1. AAA85660.1 ribosomal protein S10 Homo sapiens	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
2. XP_005553317.1 40S ribosomal protein S10 Macaca fascicularis	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
3. AAH51441.1 Ribosomal protein S10 Rattus norvegicus	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
4. AAH89323.1 Ribosomal protein S10 Mus musculus	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
5. AA02417.1 Ribosomal protein S10 Bos taurus	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
6. XP_020855249.1 40S ribosomal protein S10 Phasciarctos cinereus	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
7. XP_015675910.1 40S ribosomal protein S10 Protothophs mucrosquatus	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
8. XP_028589910.1 40S ribosomal protein S10 Podarcis muralis	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
9. XP_007063566.1 40S ribosomal protein S10 Chelonia mydas	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
10. PKK19037.1 ribosomal protein S10 Columba livia	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELVDKNVPLNHVMKAMQS
11. XP_029682140.1 40S ribosomal protein S10 isoform X1 Takifugu rubripes	MFMMLMPKKNRIAIYELFF--KEGVMVAKKDVMHLLTKH--PELADKNVPLNHVMKAMQS
12. NP_001105007.1 ribosomal protein S10 Zea mays	--ML--RNTIALRSFIRITQSTRPYVHVVEAVYYAPL--MAAKIRV
13. KAF4006697.1 ribosomal protein S10 Saccharomyces cerevisiae	--ML--RNTIALRSFIRITQSTRPYVHVVEAVYYAPL--KLPIKYGLDVAIDL
14. CAA42518.1 ribosomal protein S10 Pyrococcus woesei	--ML--RNTIALRSFIRITQSTRPYVHVVEAVYYAPL--MQKARIF
15. RVD76133.1 ribosomal protein S10 Pseudomonas koreensis	--MRITYRSHFAVCAQLIEGAKNENLVVKGPIRLPTKVLRLITTRKTPCGEGSKTWDRFGHLLFLKFYCTSNQIMAGIAYKNEKPLPDLNTEHRIIRL
16. RCN53037.1 ribosomal protein S10 Ancylostoma caninum	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
17. XP_002718159.1 40S ribosomal protein S10 Oryctolagus cuniculus	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS
18. XP_036591903.1 40S ribosomal protein S10 Trichosurus vulpecula	--MLMPKKNRIAIYELFF--KEGVMVAKKDVMHPKH--PELADKNVPLNHVMKAMQS

Можно видеть, что получившиеся выравнивания действительно отличаются. Можно видеть, что в целом, все последовательности кроме 12-16 почти идентичны (различия видны только дальше).

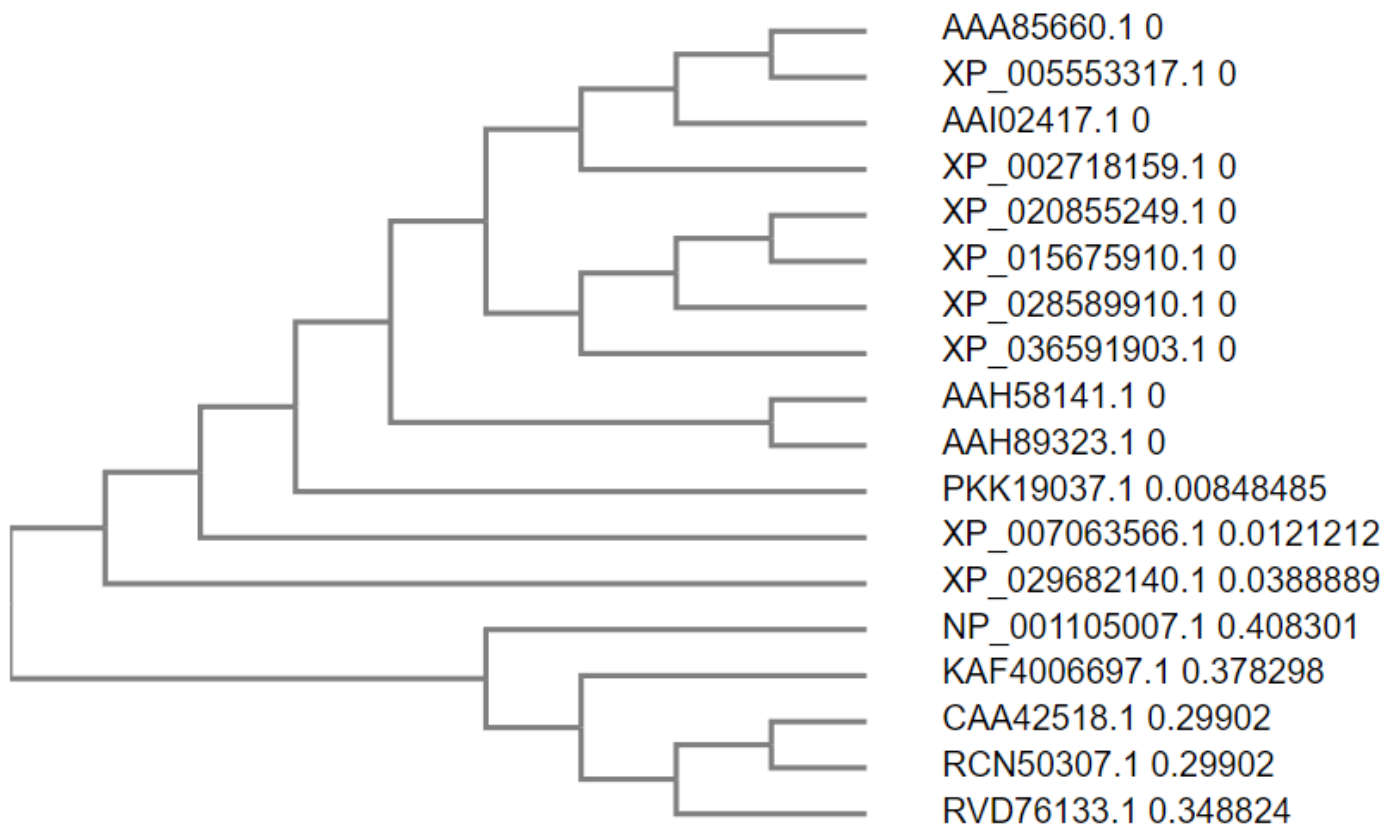
## ▼ Построение деревьев:

Теперь для каждого выравнивания будем строить выравнивания одним из 3 способов. Перед построением дерева нажимаем Phylogenetic analysis (!), только после этого выбираем bootstrap (пусть со значением 100) и строим дерево по выбранному методу.

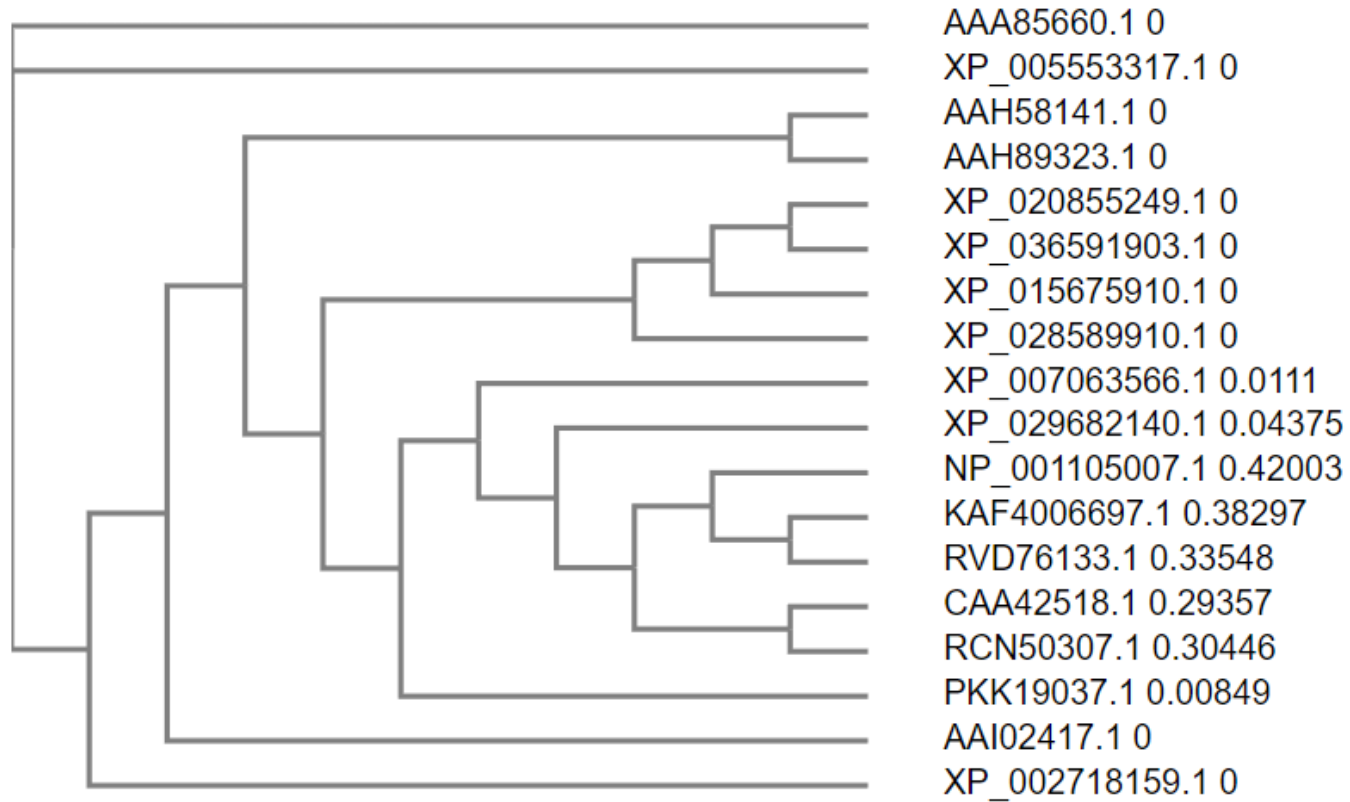
## ▼ ClustalW:

MEGA вылетает для UPGMA и NJ, обсуждал это все с Коноваловым, он сказал делать на стороннем сайте, но там нет бутстрепа, но тем не менее:

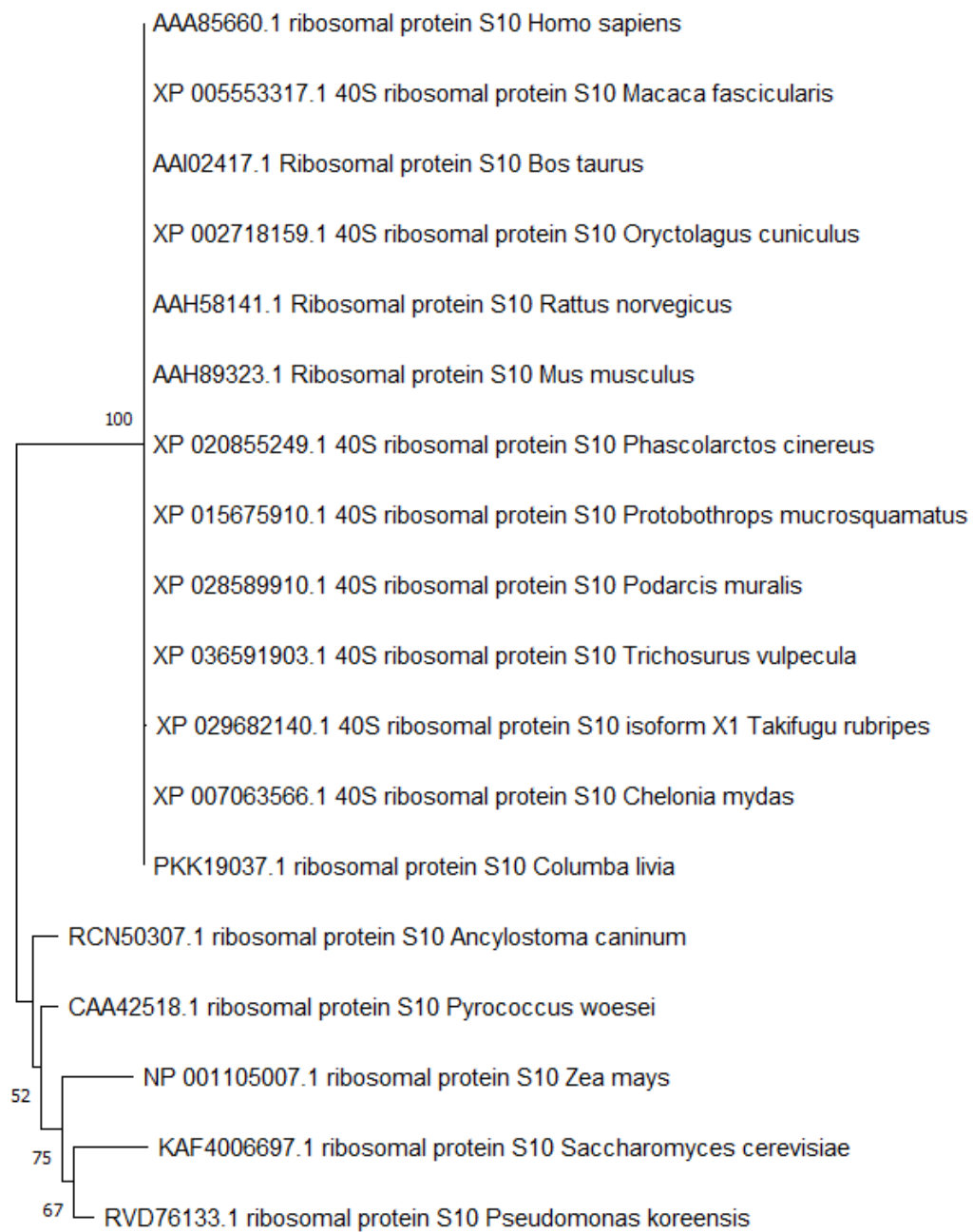
UPGMA:



Neighbor joining:



Maximum likelihood:

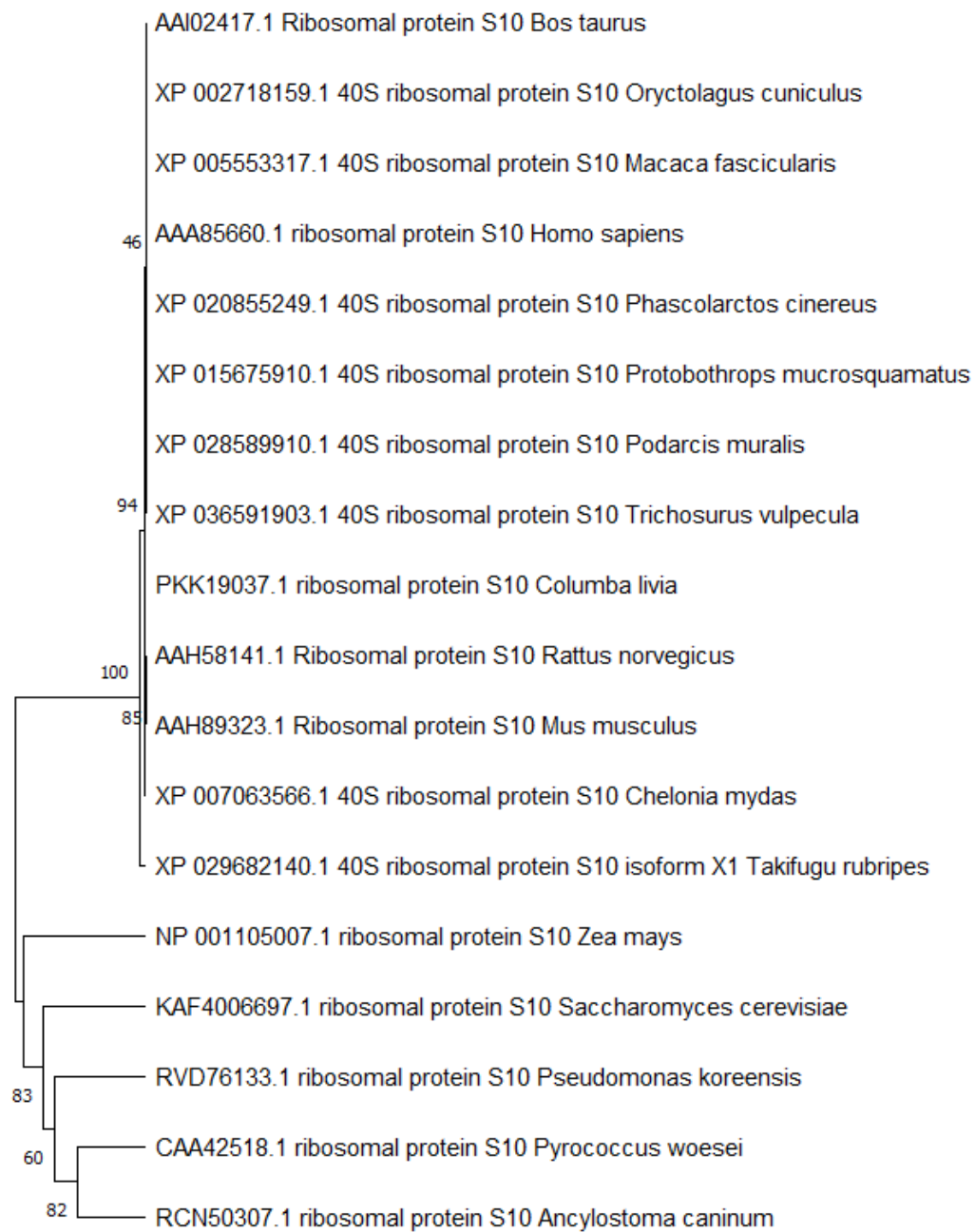


0.50

▼ MUSCLE:

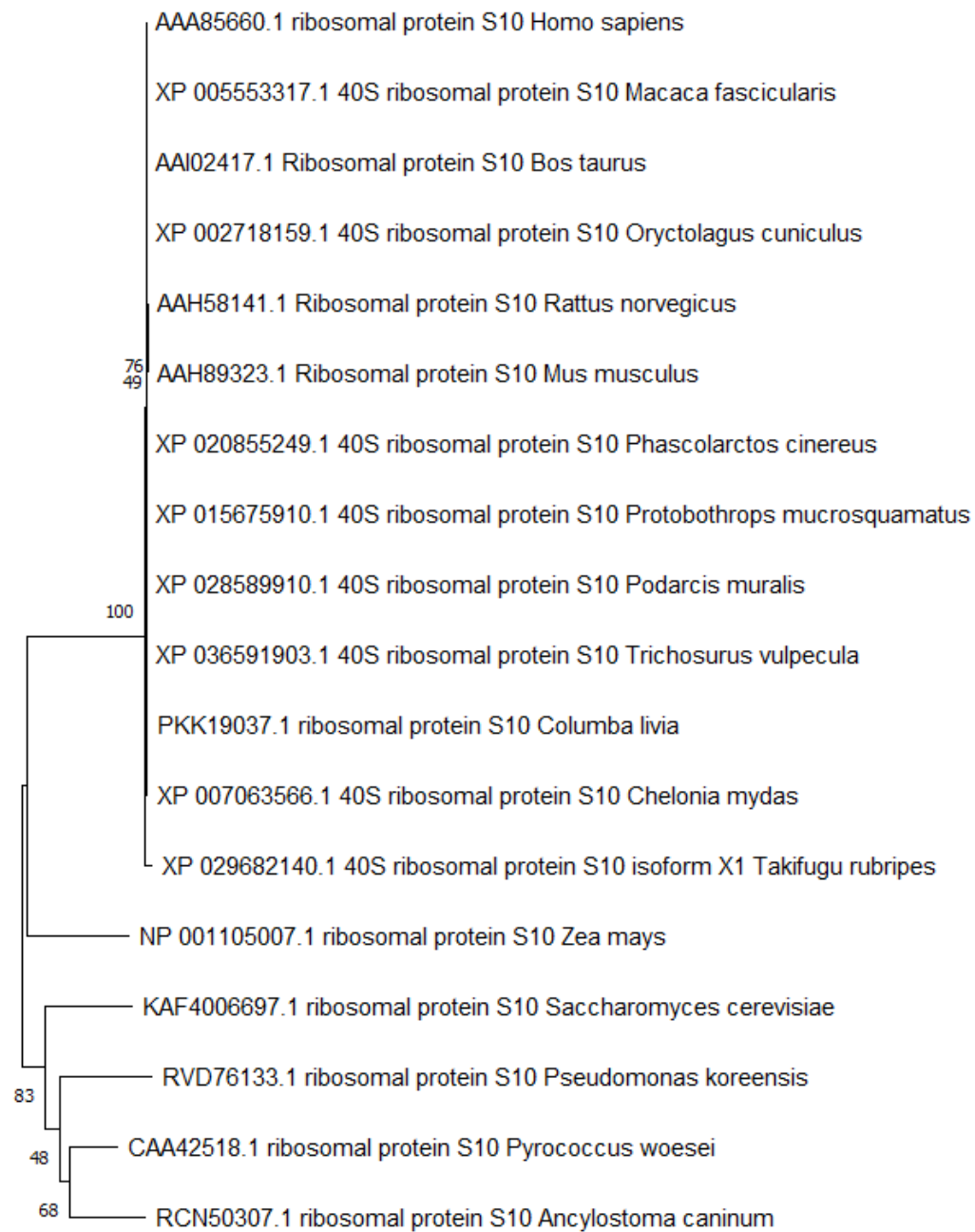
UPGMA:





0.8 0.6 0.4 0.2 0.0

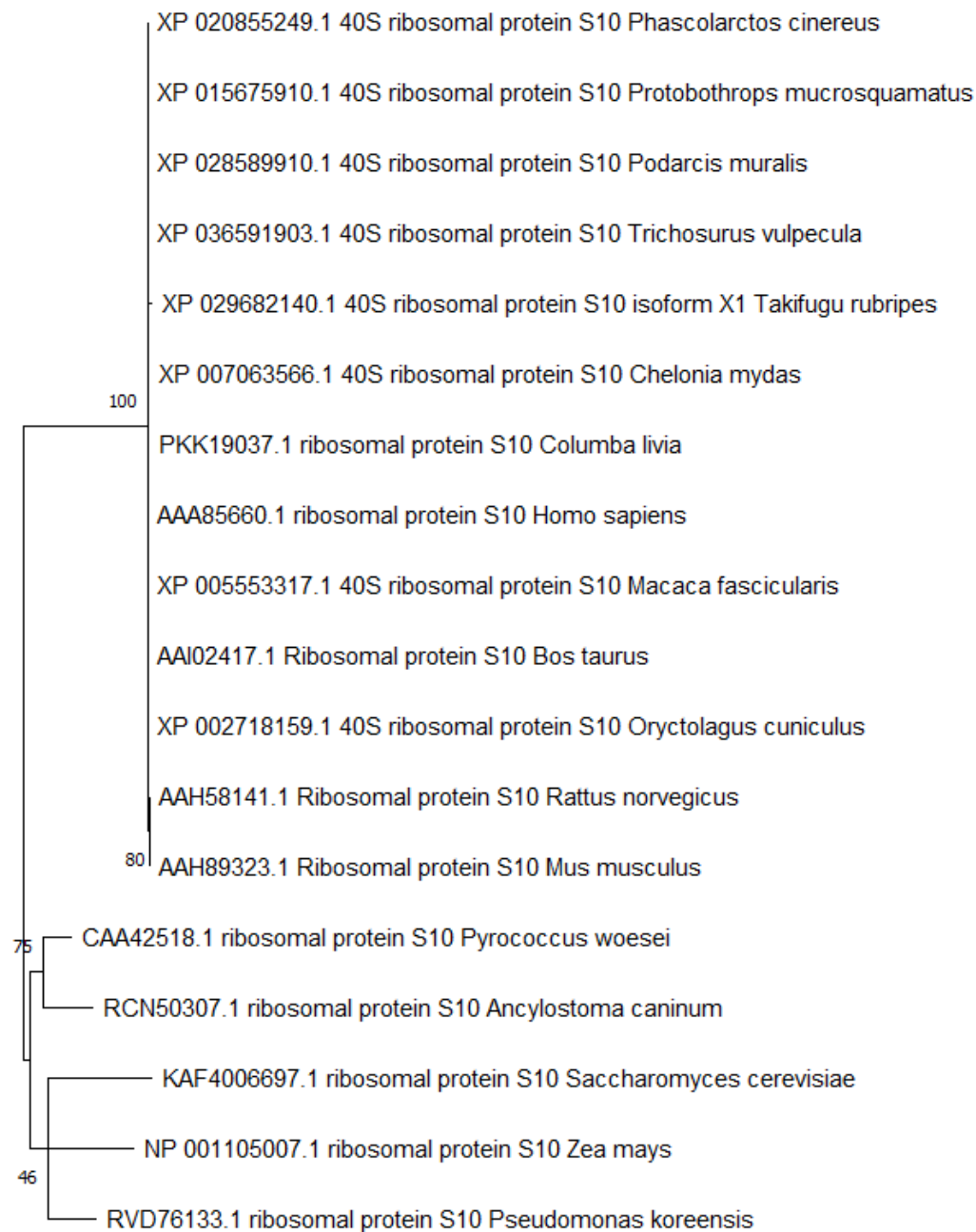
Neighbor joining:



—|—

0.20

Maximum likelihood:



Ответы на вопросы:

### 1. Какой алгоритм выравнивания лучше сработал - ClustalW или Muscle?

Ответ: Если смотреть на все выравненные последовательности для каждого метода, то Muscle получается сильно лучше. Он лучше "определил" начало и конец последовательностей, идейно целостней выглядит результат. То есть в случае с ClustalW результат выглядит так:

**Начало** - пропуск - кусок - пропуск... -кусок - пропуск,

то для Muscle видно и начало, и конец, то есть:

**Начало** - пропуск - кусок - пропуск... -кусок - пропуск - **конец**, поэтому я считаю, что он справился лучше. (**да и устроен он тоже лучше, как алгоритм и идея**)

### 2. Одинаковая ли получилась топология деревьев при построении разными методами?

Ответ: Мне немного трудно воспринимать результаты из другого приложения (для ClustalW использовали), поэтому давайте рассмотрим на примере Maximum likelihood метода.

Глобально говоря, для обоих выравниваний получилось две одинаковых по набору группы, но если смотреть более подробно, то можно видеть, что "объединялись" в группы разные последовательности, поэтому именно одинаковой назвать нельзя, но они весьма похожи.

И заметил странную деталь, что для MUSCLE выравнивания, в самом низу дерева последовательности KAF4006697.1 и RVD76133.1 как будто бы не принадлежат дереву, что очень странно, хотя, возможно, это просто баг.

### 3. Одинаковые ли получились бутстрэп-значения?

Ответ: Опять же, судя по Maximum likelihood методу, это не так. Это наблюдается в нижней половине дерева.

### 4. Совпадают ли деревья, построенные по одному гену с принятыми деревьями видов?

Ответ: Думаю, что не совсем справедливо судить по данному дереву, поскольку в верхней половине последовательности генов абсолютно идентичны (вроде бы даже все). Но, если отвечать на вопрос, то нет, не стоит, поскольку на одном уровне находится человек, мышь, копытное, птица, кролик. А, например с бактериями и археями находится ракообразное..